WPS5751

Policy Research Working Paper 5751

A Hybrid Approach to Efficiency Measurement with Empirical Illustrations from Education and Health

Adam Wagstaff
L. Choon Wang

The World Bank
Development Research Group
Human Development and Public Services Team
August 2011

Policy Research Working Paper 5751

Abstract

Inefficiency is commonplace, yet exercises aimed at improving provider performance efforts to date to measure inefficiency and use it in benchmarking exercises have not been altogether satisfactory. This paper proposes a new approach that blends the themes of Data Envelopment Analysis and the Stochastic Frontier Approach to measure overall efficiency. The hybrid approach nonparametrically estimates inefficiency by comparing actual performance with comparable real-life "best practice" on the frontier and could be useful in exercises aimed at improving provider performance. Four applications in the education and health sectors are used to illustrate the features and strengths of this hybrid approach.

This paper is a product of the Human Development and Public Services Team, Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org.
The author may be contacted at awagstaff@worldbank.org.

Produced by the Research Support Team

A Hybrid Approach to Efficiency Measurement with Empirical Illustrations from Education and Health

by

Adam Wagstaff and L. Choon Wang

Development Research Group, The World Bank, Washington DC, USA

Corresponding author and contact details: Adam Wagstaff, World Bank, 1818 H Street NW, Washington, D.C. 20433, USA. Tel. (202) 473-0566. Fax (202)-522 1153. Email: awagstaff@worldbank.org.

Keywords: Frontier, Efficiency, Cost, Education, Health.

Acknowledgements: Without wishing to implicate him in any way, we thank Martin Ravallion for helpful comments on an earlier version of the paper. The findings, interpretations and conclusions expressed in this paper are entirely those of the authors, and do not necessarily represent the views of the World Bank, its Executive Directors, or the governments of the countries they represent.

I. INTRODUCTION

In this paper we propose a new approach to measuring efficiency, and illustrate it with four applications in the education and health sectors. The paper is motivated by three interrelated ideas. The first is that inefficiency is commonplace, especially in the education and health sectors, but is not ubiquitous—some providers, some local governments, and some countries, are more efficient than others. The second idea is that having data comparing actual performance to 'best practice' could be useful in exercises aimed at improving provider performance. These might involve users holding providers accountable directly through, for example, voucher schemes. Or they might involve citizens holding politicians accountable for service delivery inefficiencies through the political process, and policymakers then holding service providers accountable through, for example, payment mechanisms that reward good performance or 'naming-and-shaming' exercises where poor performance is publicized. The third idea behind the paper is that an observation that efforts to date to measure inefficiency and use it in benchmarking exercises have not been altogether satisfactory. Some efforts focus simply on outputs or outcomes without factoring in the expenditures involved (there has been much discussion, for example, about the cross-country variation in education test scores; but it may be the case that the high achievers simply spend more), while studies that *have* sought to measure efficiency have not apparently had much impact among policymakers (cf. e.g. Burgess 2006; Hollingsworth and Street 2006); we feel this lack of impact derives from skepticism over the methods.

The method we propose and illustrate in this paper blends themes from the two efficiency measurement methodologies used to date, namely Data Envelopment Analysis (DEA) and Stochastic Frontier Analysis (SFA).[1] DEA usually focuses on estimating technical inefficiency and constructs an isoquant frontier made up of a set of piecewise linear segments joining a few data points in the input space. The frontier envelopes the data and permits the comparison of a real-life unit to a hypothetical comparator on the piecewise linear segment of the isoquant with the same input proportions. SFA, on the other hand, estimates the production or cost frontier

[1] For recent surveys of DEA and SFA in education and health, see Worthington (2001) and Hollingsworth (2008). Early published applications of SFA in education and health include Deller and Rudnicki (1993) and Wagstaff (1989) respectively. Early published applications of DEA in education and health include Bessent and Bessent (1980), Ray (1991), Sherman (1984), and Huang and Mclaughlin (1989).

using a regression model with specific functional form and distributional assumptions. The estimated frontier permits the comparison of actual output or spending to the corresponding point on the frontier to measure inefficiency. Both of these methods have pros and cons, and both share some weaknesses. Our approach tries to take the attractive features of each while trying to avoid their shared shortcomings.

We borrow two ideas from DEA and two from SFA. We focus on overall efficiency—this is straightforward in SFA but not typically done in DEA where the focus is on technical inefficiency.[2] We think policymakers will typically want to look at overall inefficiency not just technical inefficiency. On the other hand, our approach is more similar to DEA in that it envelopes the data through the use of nonparametric methods. This gets round the criticism of SFA that its results are dependent on the functional form assumed, and allows for multiple efficient units; in contrast, SFA typically produces at most a few efficient units and quite possibly just one—this will almost certainly to be the case in the classic panel-data formulation where inefficiency is modeled as a time-invariant fixed effect (Schmidt and Sickles 1984). Our approach imposes fewer assumptions than DEA, however: DEA imposes assumptions about the shape of isoquants, but we impose no assumptions about the cost curve. Finally, our approach is closer to SFA in that we make some allowance for statistical noise and measurement error. In its traditional and commonest form DEA does not allow for measurement error or statistical noise.[3] SFA, by contrast, does allow for noise and measurement error, but this comes at the expense of an arbitrary and untestable assumption that enables inefficiency to be distinguished from random shocks and measurement error.

In addition to borrowing the attractive features of DEA and SFA, we also try to avoid their shared weaknesses. Both assess the efficiency of a unit by comparing the unit's output or spending to that of a hypothetical unit rather than that of a real-life one. This use of hypothetical comparators makes the exercise untransparent to a policymaker and introduces a large element of 'make believe'. A policymaker in an inefficient country, or the head of an inefficient school board, can quite reasonably dismiss claims of inefficiency when the comparison is with a fictitious unit. And even if the policymaker or head of school district is keen to learn how to improve their performance, there is no better-performing country or better-performing school

[2] Allocative inefficiency *can* be estimated in a DEA, but has to be done explicitly, and the analyst needs to have input prices for all the inputs (Coelli 1996).

[3] Recent advances in DEA research have introduced bootstrapping in DEA to get around this (see Simar and Wilson 2000).

district to visit. In our approach, by contrast, inefficient units are compared with real-life efficient units. We can tell a policymaker of an inefficient government or a manager of an inefficient delivery unit which real-life country or unit achieves similar outcomes at a lower cost.

The second common shortcoming of DEA and SFA we try to address is their vulnerability to special pleading—policymakers or managers of poorly performing service delivery units claiming that there are legitimate factors explaining their poor performance that are ignored by the analysis. The usual response to date in the DEA and SFA literatures has been to address this issue through a two-stage approach: (1) construct or estimate a frontier to calculate efficiency scores; and (2) regress efficiency scores on factors thought to influence them. This practice has been criticized, not least because the process by which the efficiency scores is generated is ignored in the second-stage regression exercise (Burgess 2006). Instead, we build exogenous constraints into our analysis and allow different groups of units to have different frontiers. This multiple-frontier approach could, of course, be used in DEA and SFA studies too.

We illustrate our approach throughout using examples from the education and health sectors. Our examples are chosen with a view to data quality and variety in terms of level of decision-making. In both education and health, there is growing realization of the need to look beyond the number of people passing through a facility to the difference that the facility makes to the lives of the people passing through it. In the two education examples, therefore, we look at test scores to get at the quality of the education process rather than at student numbers which capture just quantity. One example comes from the level of the local government: our data come from California where school districts control the day-to-day running of public schools. The other education example relates to national education systems and makes use of data from the OECD's Program for International Student Assessment (PISA) study; while undertaken by the OECD, the geographic coverage of the study now extends well beyond the OECD countries. Our first health example also uses national data and tries to get at the efficiency of health systems. Looking at patients treated misses the quality of care, and in any case one could argue that if a health system is successful at preventing illness and injuries it ought to be reducing the number of patients requiring treatment. Measures of population health tend to be too broad-brush to be compelling measures of the outcome of a health system: many are affected by factors beyond the health system, and many causes of death are not amenable to medical care at all or only

marginally so. We therefore focus on a limited set of causes of death that are amenable to medical care, and on deaths among people under the age of 70 where for the selected conditions medical care can make a large difference. The data we use are currently available only for OECD countries. Our final example is at the facility level and concerns hospitals. A large fraction of health spending goes on hospitals, and many efficiency-enhancing efforts are directed at the hospital sector. Our data come from Vietnam, which though only a low-income country has unusually good data on its hospital sector by developing country standards. The data are not without their shortcomings, however, and this analysis especially is intended to be illustrative; with richer data on, for example, disease codes, one could do a more sophisticated efficiency analysis.

The rest of the paper is organized as follows. Section II introduces our empirical examples. Section III introduces and illustrates our hybrid approach to efficiency measurement in the simple case of a single output. Briefly, we identify efficient units through a grid-search process, identifying the least-cost unit over each output range, and then estimate a frontier nonparametrically using the output-expenditure combinations of the efficient units. We then measure the inefficiency of the inefficient units by comparing the inefficient unit's spending with the spending of the closest efficient unit on the frontier. We compare our results with those emerging from the panel-data stochastic frontier model. We obtain smaller estimates of efficiency using our hybrid approach, reflecting the fact that our frontier consists of real-life not hypothetical units. Section III extends the method to allow for multiple outputs or multiple dimensions of quality. We now identify efficient units over ranges across multiple dimensions—in the two-output case, for example, our grid search is over a square rather than a line segment. We compare empirically for each of our four examples the single- and multiple-output results. In three of the four examples, allowing for multiple outputs makes a large difference. Section IV extends the analysis further to allow for exogenous factors that constrain a unit from reaching the frontier. We illustrate our approach of different groups of units having different frontiers on the California schools dataset, using poverty as the stratifying variable. By constructing separate frontiers for school districts with small and large fractions of poor children, we allow for the possibility that the latter are constrained to making do with a lower level of 'home inputs' in the production of schooling outcomes. It turns out that allowing for separate frontiers makes less of a

difference in this example than we had expected—less than allowing for multiple dimensions of quality, for example.

II. AN INTRODUCTION TO OUR EMPIRICAL EXAMPLES

As we illustrate the methods through examples, it makes sense to introduce our examples ahead of the methods. As previously indicated, we have four examples—two from the education sector, and two from the health sector. Table 1 provides an overview of the four empirical applications, along with basic descriptive statistics.

Schools—California

It is often noted that California's school system spends relatively little per pupil by US standards, and that its students fare worse than the US national average on test scores.[4] What is less frequently mentioned is the large variation across California school districts in spending and test scores; this latter variation likely reflects the fact that the day-to-day running of California's public schools is the responsibility of school districts. A recent newspaper article by Freedberg and Doig (2011), investigative reporters with the Center for Investigative Reporting, drew attention to the large variations around the 2010 mean of $8,452 (the Pacific Unified School District spent nearly $60,000 per pupil), and noted that there is no perceptible relationship between spending and test scores. It is this intra-state variation we focus on, trying to determine which school districts are—by California's standards—efficient, and how inefficient each of the inefficient ones is.

The literature to date on the costs of California's schools has taken a somewhat different tack from ours. Imazeki (2006) estimates cost functions, allowing the structure to vary according to the concentration of schools in the school district and the fraction of pupils eligible for free or subsidized meals. No explicit allowance is made for inefficiency; for example, frontier techniques are not employed. Costrell et al. (2008) have expressed misgivings about the analysis, arguing inter alia that the results make no adequate allowance for or fail to uncover unobserved factors such as efficiency differences between school districts. Chambers *et al.* (Chambers *et al.*

[4] See, for example, The Economist April 20 2011 "A lesson in mediocrity: California's schools show how direct democracy can destroy accountability".

2006) use a panel of professional educators to assess the minimum spending necessary to deliver instructional programs for schools of varying size and demographic composition and conclude that California's schools underspend. They conclude that only 15 to 28 of the 984 school districts examined were spending at the level adequate to reach California's content and performance standards in all major subjects. The authors estimated that an additional $24.14 to $32.01 billion would have been necessary in the 2004/2005 school year to ensure the opportunity of all students to reach the state's content and performance standards. This approach begs the question of why some school districts appear to achieve much better results with similar levels of spending per pupil, and why some spend considerably more than other and yet achieve no better results.

We sourced school districts' educational outcomes and expenditure data from the California Department of Education's website.[5] Each year, the California Department of Education collects and reports information about student test scores in California Standards Tests (CST), current expense of education, enrolment, number of dropouts, number of graduates, and enrolment of English learners of roughly 980 local educational agencies (LEA) or school districts in California. For the purpose of our analysis, we focus on 313 of the school districts that control both elementary and high school levels (i.e., unified school districts) within its boundary and have data on CST scaled scores for math and reading in grade 5 and grade 7, current expenditures, enrolment, and poverty estimates consistently available between 2003/2004 and 2008/2009 academic years.[6] Californian students attending grade 2 to grade 11 take the CST in mathematics, reading, science, history, social science, and so on. Some students with disabilities take the California Alternative Performance Assessment (CAPA) and their test scores are excluded from our sample. However, students in grade 8 and beyond tend to take different CST depending on their course selection in school. We chose to focus on math and reading achievement in grade 5 and grade 7, so we have a manageable number of test outcomes that correspond to elementary school level and secondary school level.

International education systems—PISA

Much has been written about the international variation in knowledge and skills emerging from exercises such as the OECD's Program for International Student Assessment (PISA) that

[5] The website is http://www.cde.ca.gov. The current expenditure data are adjusted to June 2000 dollar values based on the Consumer Price Index (CPI) available at the Bureau of Labor Statistics website ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt.
[6] 14 of these school districts have some test score or enrollment data missing in some years because of their relatively small size.

assesses 15-year-old students in grade seven or higher every three years.[7] Four of the five top places in the 2010 report in both reading and mathematics went to Asian countries: China (Shanghai and Hong Kong SAR, China), S Korea, and Singapore. By contrast, the US ranked 11th on reading and 26th on mathematics. Discussions of these results, such as OECD (2010), typically focus simply on scores, without factoring in the amounts that countries spend on education. It is perfectly possible in principle that the countries achieving the best outcomes are those that spend the most, and that those who do less well spend relatively little. The more interesting—but rarely asked—question from a policy perspective is how countries vary in their success at translating resources into learning outcomes.[8]

We explore this issue using data on knowledge and skills in mathematics, reading and science from the PISA exercise, and secondary education spending data from the World Bank's World Development Indicators (WDI). We have an unbalanced panel comprising 118 sets of test scores and secondary education expenditure per pupil for 49 countries for some or all the years 2000, 2003 and 2006.[9] Because information about expenditure per student is missing in the WDI for some countries in some years, 25 of the 118 observations are interpolated or extrapolated. Although in each assessment of PISA, one of the three areas (science, reading and mathematics) is selected as the major domain and is given greater emphasis, while the two minor domains are assessed less thoroughly, differences in test scores across countries and over time are comparable because the tests are linked assessments and the scores are scaled.[10]

International health systems—OECD

It is well known that countries vary considerably in their per capita spending on health care, with the US being among the largest spenders per capita. Inevitably, there has been and continues to be a lot of debate on the question of whether the high spenders (especially the US) achieve sufficiently better results to warrant the extra spending (cf. e.g. Anderson and Frogner 2008), and more generally on how to measure health sector results in international comparisons

[7] See http://www.pisa.oecd.org/ for further details of the PISA program. For studies examining the cross-country differences in PISA, see Dobert, Klieme, and Sroka (2004), Ammermueller (2007), and Fuchs and Woessmann (2007).
[8] Afonso and St Aubyn (2005) is an exception. They use DEA to analyze efficiency in education spending in OECD countries using the PISA data.
[9] We have not included data for 2009 because most countries do not have expenditure data available in the WDI database. The secondary education expenditure per student is generated by multiplying purchasing power parity GDP per capita (2005 constant value) with expenditure per student in secondary education as a percentage of GDP per capita.
[10] See PISA 2003 Technical Report (Adams 2005) for details.

of health system performance (see e.g. Hakkinen and Joumard 2007). Throughput measures—such as inpatient admissions, and ambulatory care visits—are just that, and need not necessarily capture the ultimate outcome of interest, namely better health. But health indicators such as mortality (or life expectancy) and disability reflect many factors beyond the control of the health sector: deaths from road accidents likely reflect road safety improvements more than health spending, and many deaths are from causes that are still not amenable to medical intervention. Arguably a more compelling approach when assessing the efficiency of a country's health spending is to link health spending to deaths from conditions that *are* amenable to medical care (Nolte and McKee 2008). This is what we do in this study.

Our data come from the OECD's Health Data. We have data for 29 countries over the period 1960-2005. Not all countries have data for every year. Health spending is defined as total health spending (i.e. public plus private) measured in 2000 prices in international dollars. Mortality is measured through potential years of life lost (PYLL) among people below the age of 70 who die from nine causes of death that are 'amenable' to medical care, i.e. causes where timely and effective medical care can result in a death being avoided. We select the conditions from Nolte and McKee (2008) who identify a longer list of conditions but only nine are included in the OECD PYLL database. Deaths among older age groups are excluded by the OECD on the grounds that they are less easily amenable to medical care. We aggregate some of the conditions so we are left with a more manageable six PYLL 'outputs'.[11] Our measure of health sector output—while preferable to a throughput measure and better than all-cause mortality among all age groups—is not without its limitations, of course. It focuses on length of life rather than quality of life and does not capture success in reducing mortality among the over-70s from conditions that are amenable to medical care. Countries that disproportionately target spending at the over-70s, or at patients whose length of life cannot be extended but whose quality of life can be improved will appear in our analysis as inefficient.

Hospitals—Vietnam

Hospitals absorb the bulk of health spending in most countries, and there has been much discussion of the scope for lowering health spending by reducing their inefficiency.

[11] We aggregated PYLL's from three types of cancer (colon, breast and cervical) into an aggregate cancer PYLL, and aggregated PYLL's from pregnancies/deliveries and perinatal causes into an aggregate maternal and child health (MCH)) PYLL. The remaining four amenable causes were diabetes, ischemic heart disease, cerebrovascular disease, and influenza/pneumonia.

Unsurprisingly some of the first applications of DEA and SFA in health were on the hospital sector (e.g., (Wagstaff 1989) and (Ray 1991)), and there has been a good deal of work undertaken since then: EconLit contains 56 publications with "hospital" and "frontier" in the abstract). According to Hollingsworth and Street (2006), however, this work has had a relatively modest impact on policymakers.

Our data are from Vietnam's official public hospital inventory, the same dataset used by Weaver and Deolalikar (2004) in their study of economies of scale and scope in Vietnamese hospitals. By the standards of low- and middle-income countries, this is an unusually good dataset. However, it does lack detailed information on patients treated, distinguishing only between inpatients, surgery cases and outpatients and not between different departments, let alone different diseases and treatments. In what follows we have included only district hospitals that have between 50 and 500 beds. We have excluded central hospitals run directly by the health ministry, and level-1 and level-2 hospitals (more complex hospitals). Our sample consists of 795 hospitals. Our data are for three years: 1998, 1999 and 2000. The expenditure data cover recurrent costs.

III. **THE SINGLE OUTPUT CASE**

We start with the simplest case, where we have just one output, or one dimension of quality. (We allow for multiple outputs in the next section.)

Methods

We assume labor and nonlabor inputs are combined to produce an output y at a cost C. Costs can exceed their feasible minimum because the input bundle used does not yield the maximum possible output (technical inefficiency), or because inputs are used in the wrong proportions given their prices and marginal products (allocative inefficiency), or both. We do not try to disentangle the two, instead presenting an estimate of overall inefficiency.

Suppose we have data from multiple service-delivery units. We can then generate a scatterplot of C (or average cost) against y—the space of the standard total (or average) cost curve chart in a microeconomics textbook. In services like education and health, it is important

to allow for quality and not just focus on 'outputs' such as enrollments or cases treated. We can allow for quality by graphing average cost per person, C/y, against quality, q. For example, y might be students enrolled, C/y cost per student, and q the average test score.[12]

In the first stage of our analysis, we identify a group of efficient service delivery units (or, in the case of panel data, efficient service delivery units *at a point in time*), defined as those that have the smallest (total or average) expenditure for each level of output, or the smallest expenditure per student or patient treated for each level of quality. Because there will be relatively few units that have *exactly* the same output (or quality), we work with output (or quality) *ranges*. We define a caliper of size c, and move the caliper along the y (or q) axis in steps of size $s \leq c$. In this case where y (or q) is a scalar, the caliper is a line of length c which gets moved up the y (or q) axis up to the maximum value of the outcome in steps of s. In each step, the unit with the smallest expenditure within the caliper is identified and labeled an 'efficient unit'. Next we create a (stochastic) frontier by running a nonparametric Lowess smoother through the datapoints of these efficient units, and defining the frontier as the predicted cost for each efficient unit. The grid-search process thus identifies efficient units, and the smoothing process produces the frontier, with all efficient units being moved to the frontier.

In the second stage, we compute the inefficiency of inefficient delivery units by matching each unit off the frontier with the closest unit on the frontier in terms on the outcome y; the unit's inefficiency is the difference between its expenditure and the expenditure of the closest match on the frontier.[13] In contrast to both DEA and SFA where units are compared with hypothetical units, the matched unit for each inefficient unit in our approach is a real-life service delivery unit, not a hypothetical point on the frontier. We see this as a strength of our approach; all are units that have actually managed to produce (close to) output y at a cost C, not ones that ought to have been capable of doing so.

Two points are worth clarifying at this stage. First, the first stage may leave some inefficient units below the smoothed frontier. These are units that emerge with somewhat higher

[12] For such a graph to be justified, the underlying two-product cost function would have to have the form $C(y,q) = y \cdot c(q)$, giving $C(y,q)/y = c(q)$. Crampes and Hollander (1995) use such a cost function, but do not explore its properties. For the most part, the properties are fairly innocuous. The extent of ray economies (the effect on cost of doubling both y and q) depends on the shape of $c(q)$, with $c'(q)>0$ implying ray diseconomies and $c'(q)<0$ implying ray diseconomies. There are economies of scale with respect to quality if $c'(q)<c(q)$. The only odd feature of the cost function is that it implies that average incremental costs for quantity always exceeds the marginal cost for quantity.

[13] We use the Stata module PSMATCH2 (Leuven and Sianesi 2003) to do the matching.

costs than the least-cost unit over a specific output range. Because the latter have their actual cost replaced by the predicted cost from the smoother, some units will be moved up to the frontier, and hence may find themselves with 'expected' costs that are higher than the actual costs of inefficient units, albeit ones that have costs that are close to the frontier costs. We reclassify such units as efficient units ex post, by setting their inefficiency scores to zero in the second stage. Second, our frontier is inevitably dependent on the size of the caliper *c* and the step *s*, as well as potentially the bandwidth of the Lowess smoother. Our empirical analysis therefore inevitably entails some sensitivity analysis to see how sensitive inefficiency scores (and the rankings of service delivery units) are to the choices of *c* and *s*.[14]

The use of the Lowess smoother allows for a flexible representation of the frontier, while the grid search ensures that for each output (or quality) range we base the frontier on a real-life unit. Our approach blends themes from DEA and SFA. Like DEA, our approach identifies efficient units through an envelopment process using a subset of datapoints; this contrast with SFA which uses all datapoints to establish the frontier. Of course, we work in a different space from DEA, which operates in input space. The fact that we replace actual costs with predicted costs is reminiscent of SFA; our use of a smoother accepts that there is some randomness associated with realized costs at each output level. Insofar as we have repeated observations over time on a given service-delivery unit (in our empirical examples we do), we can average inefficiency scores over time to allow for randomness in the realized expenditures of inefficient units. Unlike SFA we make no assumption about the functional form of the cost function, and we use only a subset of datapoints to arrive at the frontier.

Empirical examples

Figure 1 illustrates the idea in each of our four cases. The first three allow for quality by graphing expenditure per person against a quality measure—test scores in the two education examples, and in the case of national health systems the aggregated potential years of life lost occurring before the age of 70 through deaths from six causes amenable to medical care. The large dots correspond to units that emerge as efficient in our grid search, while the smaller dots correspond to the units that do not emerge as efficient through this process. The hybrid frontier in each example is constructed by passing a Lowess smoother through the efficient points.

[14] We do not report results for different bandwidths of the Lowess smoother.

The charts illustrate the need to look at both spending and outcomes, particularly in education where for the most part higher quality entails on average higher spending per pupil. The PISA data reveal, for example, that Uruguay fares relatively badly in terms of test scores, but also spends relatively little; in fact, it ends up on our frontier. Sweden, by contrast, does much better in terms of test scores, but spends considerably more and ends up some way above our frontier. Finland does well both in terms of its test score results and its efficiency in translating resources into outcomes; it ends up on or close to the frontier depending on the year.

Figure 2 shows the least efficient units in each of the four studies. These are all units that achieve similar outcomes to other units but at a considerably higher cost. Interestingly six of the seven least efficient countries in the education rankings are among the least efficient seven in the health system rankings. For the California schools example, the majority of the least efficient school districts listed, such as Shoreline Unified and San Pasqual Valley Unified, tend to be tiny districts located in sparsely populated remote areas, a feature similar to the Pacific Unified that Freedberg and Doig (2011) noted.

Figure 3 compares our hybrid approach with the stochastic frontier model (SFM) where inefficiency is assumed to be half-normal and time-invariant (in all four examples, we have panel data) and the relationship between expenditure and output or quality is assumed to be double logarithmic. (Interestingly, in all these applications the cross-section SFM concluded that there was no cross-sample variation in inefficiency.) The charts show the hybrid frontier and the actual expenditure-output combinations along with the deterministic section of the SFM and the efficient expenditure levels emerging from the SFM. These differ from the expenditure corresponding to the deterministic section of the frontier because the SFM approach allows for statistical noise and random shocks. In each example, the deterministic section of the SFM and most of the efficient spending levels lie well below our hybrid frontier. The SFM typically finds one or at most a few units to be efficient; the inefficiency components of the residuals of the rest are adjusted accordingly. By contrast, in our approach, we find an efficient unit *for each output range*. The fact that the SFM imposes a specific functional form also results in differences between the efficiency scores emerging from the two approaches; these are particularly large in the case of the California schools example, but are also evident in the PISA and OECD examples particularly at the upper quality levels. Of course a different functional form in the SFM could

reduce these discrepancies, but the nonparametric element in our hybrid approach avoids the need to take decisions about the functional form of the deterministic component of the SFM.

The first column Table 2 for each example shows the estimated mean inefficiency in each of the four exercises for our hybrid approach. Also reported are standard errors for these estimates. These are not the standard errors produced through the matching routine, since this does not take into account the fact that the frontier is itself subject to sampling variability. Rather we obtain the standard errors in Table 2 through bootstrapping. We bootstrap the entire process: the grid search for efficient units; the Lowess smoothing to obtain the frontier; and finally the matching of inefficient to efficient units to measure inefficiency. Also reported are the calipers and steps, and the number and fraction of datapoints on the frontier. Inevitably in all efficiency measurement exercises, the inefficiency estimates depend on the sample. The higher inefficiency estimates in the PISA exercise than in the health system exercise reflects the fact the PISA exercise, while undertaken by the OECD, includes some non-OECD countries, some of which spend less on education and achieve good test scores; these end up on the frontier, raising the average inefficiency compared to the health system exercise where these countries are absent from the dataset.

How sensitive are our results to the choice of caliper and step? We explore this by varying the step and caliper and seeing how the average inefficiency changes, and how the ranking of service delivery units (in a given year) changes. Figure 4 shows the sensitivity of the inefficiency estimates to the choice of caliper and step. As we raise the caliper, the number on the frontier falls, and average inefficiency increases. However, even for a small caliper, and hence a larger number of frontier units, inefficiency is appreciable in all four examples. Changing the step for a given caliper has a smaller effect in the examples where we have a relatively large number of service delivery units at each point in time—California school districts, and Vietnamese hospitals. The surface in Figure 4 is much "choppier" for the two country-level examples, particularly for small calipers. Figure 5 shows how sensitive the rankings of service delivery units are to the choice of caliper and step. We compare the rank for each caliper and step combination to the rank in the distribution underlying Table 2—our basic results. The rankings are remarkably robust to the choice of caliper and step; the most fragile rankings are the PISA results, where the rank correlations hover around 0.8. This reflects the small size of the dataset, and the considerable heterogeneity among the countries included.

IV. MULTIPLE OUTPUTS AND MULTIPLE QUALITY DIMENSIONS

In many applications, we may want to allow for multiple outputs, or for multiple dimensions of quality. In the two education examples we have averaged tests scores in each area; in reality, it is likely that teaching mathematics and reading (say) may entail different costs. This is even more likely to be the case in the health systems example, where we have aggregated different 'amenable' causes of death. In the hospital example, we have aggregated outpatient, inpatient, and surgery cases; in reality the three may have different cost structures. In this section, we generalize our hybrid approach to efficiency measurement to allow for multiple outputs and multiple dimensions of quality.

Methods

Consider the two-output case. We have an output vector y containing two outputs, y_1 and y_2. In this case, the caliper becomes a square of length and width c, which gets moved in steps of s along both dimensions of the plane defined by the minimum and maximum values of the two outcomes. Within each square, we identify the least-cost unit; a multidimensional Lowess smoother is then run through the vector (C,y_1,y_2) for the efficient units which creates the frontier surface. Each unit that is above the frontier surface has costs that exceed the lowest recorded in the output 'square' in question. We then match these inefficient units to units on the frontier using the Mahalanobis (1936) metric; our outputs do not need to be measured on the same metric. This two-stage process can be generalized to three or more outputs. Obviously the process becomes more involved the more outputs that are included in the exercise.

The approach can be generalized to handle cases where there are multiple dimensions to quality. For example, y might be students and we might want to explicitly allow for quality across two dimensions: q_1 might be the average test score in Mathematics and q_2 might be the average test score in Reading. In this case we have a three-dimensional chart of C/y against q_1 and q_2. The approach to efficiency measurement proceeds as in the multiple-output case. This can be generalized to more than two dimensions of quality.

Empirical examples

In each of the four examples, we re-estimated the inefficiency scores allowing for multiple quality dimensions or multiple outputs. In the education examples, instead of averaging test scores across subjects we allowed for four quality dimensions in the California case and three in the PISA case. In the health systems example, we also allowed for multiple quality dimensions, allowing for six separate amenable causes of mortality. In the hospital example, we allowed for three outputs, outpatient cases, inpatient cases and surgery cases.

Figure 6 compares the distributions of inefficiency scores for the single- and multiple-output exercises. In the education examples, the switch to multiple outputs/dimensions results in an appreciable leftward shift of the inefficiency distribution; many units that looked inefficient when outputs or quality are reduced to one dimension look less inefficient when we allow for different cost structures across different outputs or quality dimensions. This is also evident comparing the two columns in Table 2 for each study. In the California case, the shift is quite dramatic, reducing mean inefficiency from $3,997 to $1,931. The top two maps in Figure 9 show this graphically: the right-hand map, which allows for multiple outputs, is considerably lighter in shade than the left-hand one, which is based on average test scores across all subjects. Figure 7 sheds light on how much reranking goes on in the move from the single-outcome to multiple-outcome exercises. In the California case, the scatter plot is quite busy, implying a fair amount of reranking. The Vietnam scatter is much tidier.

V. ALLOWING FOR EXOGENOUS CONSTRAINTS ON EFFICIENCY

A common response from policymakers and managers when confronted with estimated efficiency scores from DEA and SFA is that the analyst has made no allowance for factors that are beyond the policymaker's or manager's control that prevent them from operating on the frontier. A school may be operating in a catchment area with a very poor population and may not receive additional funding to compensate. A health facility may be operating in a sparsely populated or topologically challenging area, making outreach difficult.

In DEA and SFA, the response has typically been to 'explain' inefficiency scores through regression analysis to such constraints and other influences on efficiency that are under the control of the decision-making unit, although surprisingly, such analysis does not, as one might

expect, adjust the efficiency scores to take into account the effects of the exogenous constraints. This ex post regression approach has in any case been criticized on the grounds that it makes no allowance for the way the scores are generated (cf. e.g. Burgess 2006).

Methods

A more fundamental objection to the post-hoc analysis of the influence of exogenous constraints on efficiency scores is that such constraints ought to be allowed for *during* the computation of the efficiency scores. In our hybrid approach we do this by allowing different groups of units (defined in terms of constraints) to have different frontiers.[15] This allows for considerable flexibility—much more than would be obtained than by, for example, including a constraint variable in a stochastic frontier. Our approach allows the location and shape of the frontier to vary across the groups. For each group, having generated the frontier, we compute an inefficiency score relative to the relevant frontier.

Empirical example

We illustrate the idea of allowing for exogenous constraints using the California schools data. We stratify school districts by poverty, distinguishing between districts in the bottom half and top half of the poverty headcount distribution. The rationale is that pupils from more affluent families may face more favorable home environments than pupils from poor families: the time available for homework may be greater; the degree of parental input and oversight may be greater; the pupils may come to school better nourished; and so on. School inputs may be comparable, but home inputs that are beyond the control of the school may be smaller in schools in poorer catchment areas.

The school district estimates of school age students (aged 5 to 17) living in households under poverty were sourced from the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE).[16] According to the U.S. Census Bureau's website, the estimates were computed using data from administrative records, intercensal population estimates, and the decennial census with direct estimates from the American Community Survey.

[15] This principle could also be applied to other approaches to efficiency measurement, of course. One could, for example, estimate separate stochastic frontiers for different groups.
[16] The data are available at http://www.census.gov/did/www/saipe/data/schools/index.html

The top-left panel of Figure 8 imposes a single frontier but colors the low-poverty school districts lighter (in green) and the high-poverty ones darker (in red). The high-poverty districts have worse test scores in absolute terms and at a given level of spending. The top-right panel shows that a single frontier produces higher inefficiency estimates for the high-poverty districts than for the low-poverty districts. There is a prima facie case therefore that separate frontiers might be warranted. The bottom-left panel allows for different frontiers, with the frontier for the high-poverty districts being (mostly) similarly shaped but somewhat to the left of that of the low-poverty districts. The bottom-right panel shows that separate frontiers yield a much more similar inefficiency distribution between the poor and less poor school districts than a single frontier.

Comparing the top and bottom maps of each pair in Figure 9 shows graphically the effect on the inefficiency distribution of allowing poor and less poor districts to have different frontiers: both the one-output and multiple-output maps become lighter in shade as we move from the top where no allowance is made for poverty to the bottom where poor and less poor school districts are allowed to have separate frontiers. It is noteworthy, however, that the lightening of shade is less than is the case when one moves rightwards in Figure 9 from the single-output model to the multi-output model. The implication is allowing for different frontiers across different outputs is already sufficient to take into account the differing circumstances of the poorer and less poor school districts.

VI. **CONCLUSIONS**

In this paper, we propose a hybrid approach to measuring efficiency that blends the themes from the Data Envelopment Analysis (DEA) and Stochastic Frontier Analysis (SFA) and illustrate different aspects of it with four empirical applications. We conclude the paper by recapping the strengths of our hybrid approach, summarizing the steps to be followed in implementing it, and highlighting the key aspects of the methods demonstrated in our empirical illustrations.

Our approach has several strengths. First, it focuses on overall efficiency and does not require data on input prices. Second, the efficiency of a unit is assessed against a comparable real-life unit instead of some hypothetical comparators. Third, the approach is nonparametric, imposes few distributional and functional form assumptions, and allows for statistical noise in

the process of constructing the frontier. Fourth, it can flexibly accounts for exogenous constraints that lead to poor performance.

The major steps involve in the hybrid approach are:

1. Define the size of caliper and step;
2. Use a grid search with a caliper and step to identify within each grid the unit with the smallest expenditure, and define as efficient the units identified through this process;
3. Create a stochastic frontier by running a nonparametric Lowess smoother through the efficient units;
4. Replace actual expenditures with predicted expenditures for efficient units;
5. Match each inefficient unit with the closest efficient unit;
6. Compute inefficiency scores;
7. Replace inefficiency scores of inefficient units *below* the efficiency frontier to zeroes;
8. Vary the choice of caliper and step to assess the sensitivity of rankings.

If exogenous constraints on efficiency are to be allowed, then the steps are:

1. Define several ranges of values within a constraint and create a cell corresponding to each range of values within a constraint;
2. Apply steps 1 to 8 above within each cell to generate inefficiency scores;
3. Undertake a sensitivity analysis of rank correlation with respect to choices of caliper and step.

The empirical illustrations demonstrate several interesting aspects of the hybrid approach. First, we show that the hybrid approach envelopes the data better than SFA and produces smaller (and, we think, more realistic) inefficiency scores. Second, inefficiency scores tend to be lower in multiple-output cases than in single-output cases. Third, although the choice of caliper and step influences inefficiency scores, the rankings of inefficiency scores are remarkably robust, especially when the sample size of the dataset is relatively large. Finally, the California school districts example illustrates that allowing for exogenous constraints reduces inefficiency scores, though the reduction in inefficiency scores is larger by switching from the single-output case to the multiple-output case.

Table 1: Data for the four empirical illustrations

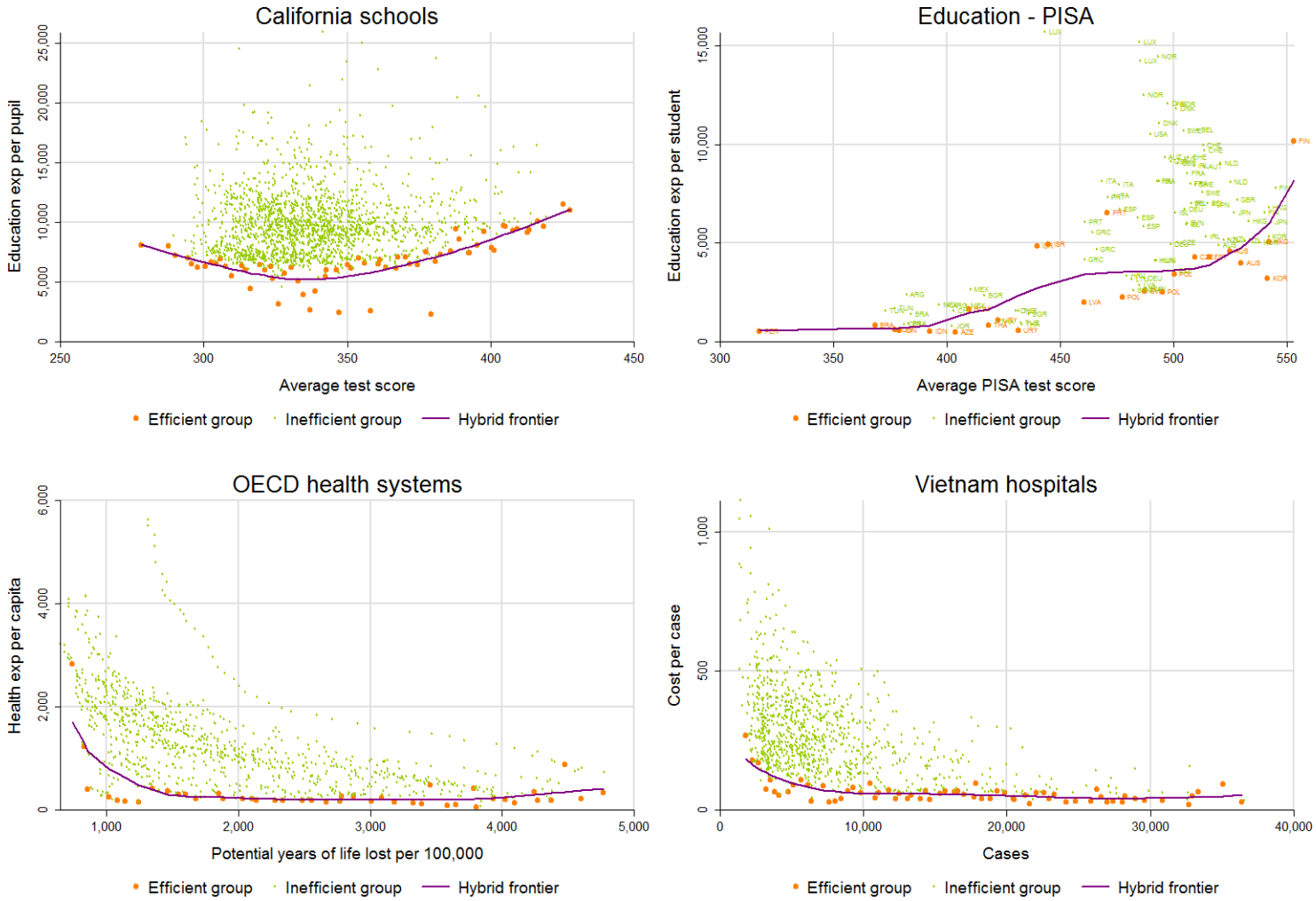
	California school districts	OECD Program for International Student Assessment (PISA)	OECD health systems	Vietnam district hospitals
Sample coverage	313 unified school districts in California.	49 countries.	29 OECD countries.	795 district hospitals in Vietnam with 50-500 beds.
Time period	2003/04-2008/09	2000, 2003, 2006	1960-2005	1998, 1999, 2000
Sample size	1878	118	945	1039
Expenditure measures	Current expenditures in June 2000 prices, per pupil	Secondary education expenditure in international dollars at 2005 prices, per secondary pupil	Total health spending in international dollars at 2000 prices, per capita	Recurrent costs of hospital, per case
Mean and standard deviations of expenditures	9742.44 (2662.77)	5492.25 (3637.27)	1294.55 (891.70)	262.83 (158.36)
Outcome measures	California Standards Test (CST) scores of grade 5 and grade 7 pupils in math and science. Averaged for one-outcome exercise.	Average PISA test scores in math, reading, and science of 15-year-old students in grade 7 or higher. Averaged for one-outcome exercise.	Potential years of life lost (PYLL) per 100,000 population through deaths occurring before the age of 70 from six causes of death considered to be amenable to medical care. Aggregated for one-outcome exercise	Outpatient and inpatient cases. Aggregated for one-outcome exercise
Outcome [mean; std. dev.]	Grade 5's math [356.25; 33.96] Grade 5's reading [346.58; 22.20] Grade 7's math [339.05; 24.81] Grade 7's reading [345.54; 22.74] Average outcome [346.86; 24.50]	Math [475.85; 56.61] Reading [473.34; 47.45] Science [479.68; 48.30] Average outcome [476.33; 49.86]	PYLL-agg. cancer [378.20; 106.50] PYLL-MCH [584.18; 444.59] PYLL-diabetes [52.67; 36.84] PYLL-heart disease [530.55; 306.48] PYLL-cerebrovascular [228.62; 134.89] PYLL-Flu/pneumonia [149.29; 178.54] Aggregate outcome [1923.51; 924.54]	Inpatient cases [4914.95; 3021.12] Outpatient cases [2546.67; 4665.88] Surgery cases [423.23; 501.74] Aggregate outcome [7884.85; 6065.08]
Data sources	California Department of Education Data	PISA and WDI Databases	OECD Health Data	Vietnam annual hospital inventory exercise

Table 2: Results from hybrid efficiency measurement method

	California school districts		OECD Program for International Student Assessment (PISA)		OECD health systems		Vietnam district hospitals	
	Single	Multi-outcome	Single	Multi-outcome	Single	Multi-outcome	Single	Multi-outcome
Inefficiency	3997	1931	2318	1721	868	1056	172	172
t-stat	27.29	6.19	8.26	2.95	16.38	20.94	18.96	15.88
Caliper	9.0	15.0	10.0	12.0	9.0	5.0	7.5	4.0
Step	9.0	20.0	10.0	16.5	9.0	2.5	7.5	4.0
No. of datapoints on frontier	66	67	25	19	44	50	66	65
% on frontier	3.5%	3.6%	21.2%	16.0%	4.7%	5.3%	6.4%	6.3%

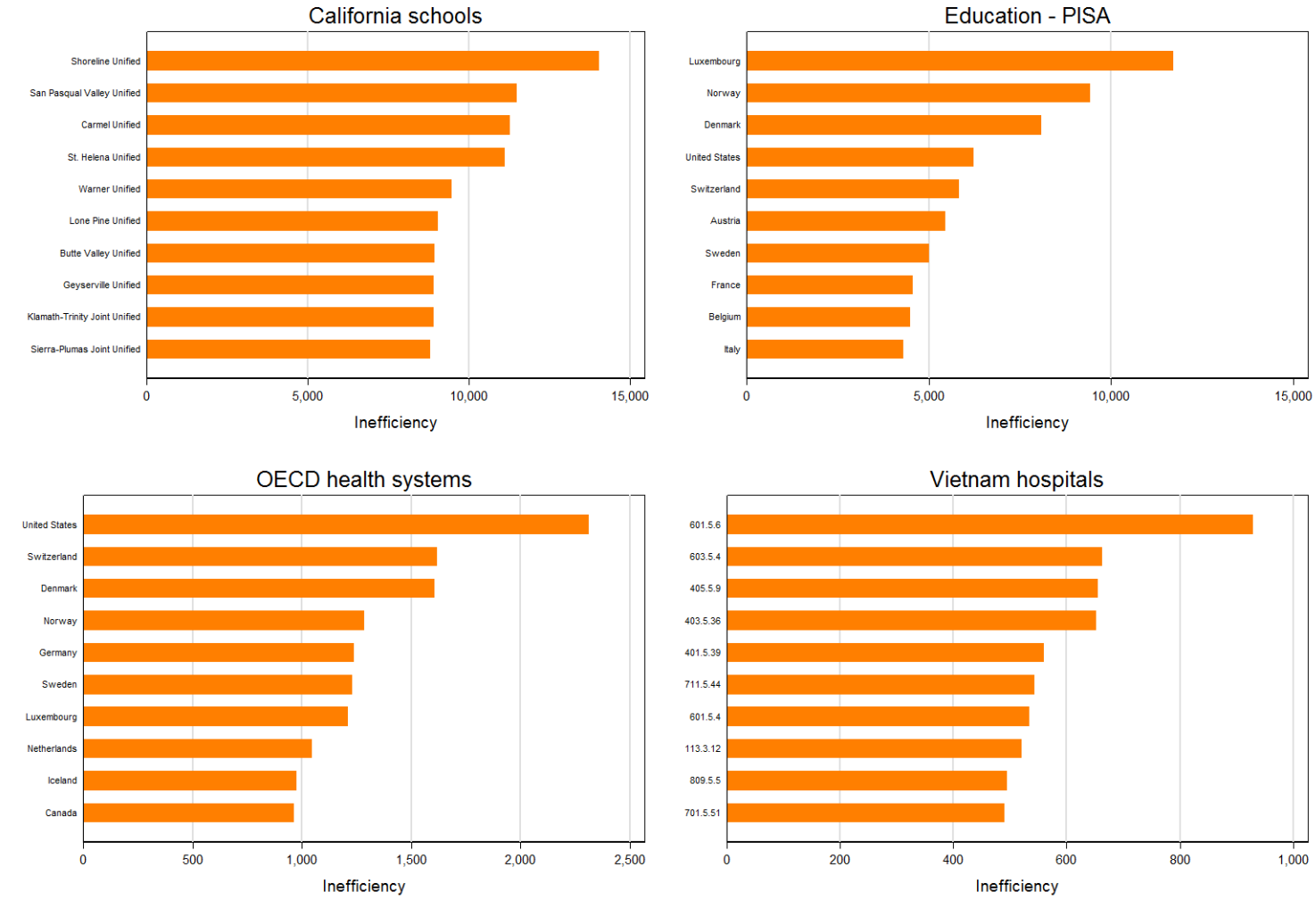
Notes: t-stat is derived from the standard error obtained via bootstrapping with 250 replications. Actual caliper and step used are 1/100 of the values shown in the case of the single-outcome exercise, and 1/10 of the value shown in the case of the multi-outcome exercise. The caliper and step are both defined in terms of standard deviation of the outcome variable. The calipers and steps in the single- and multiple-outcome cases are chosen to allow for similar number of datapoints on the frontier.

Figure 1: Illustration of hybrid efficiency method



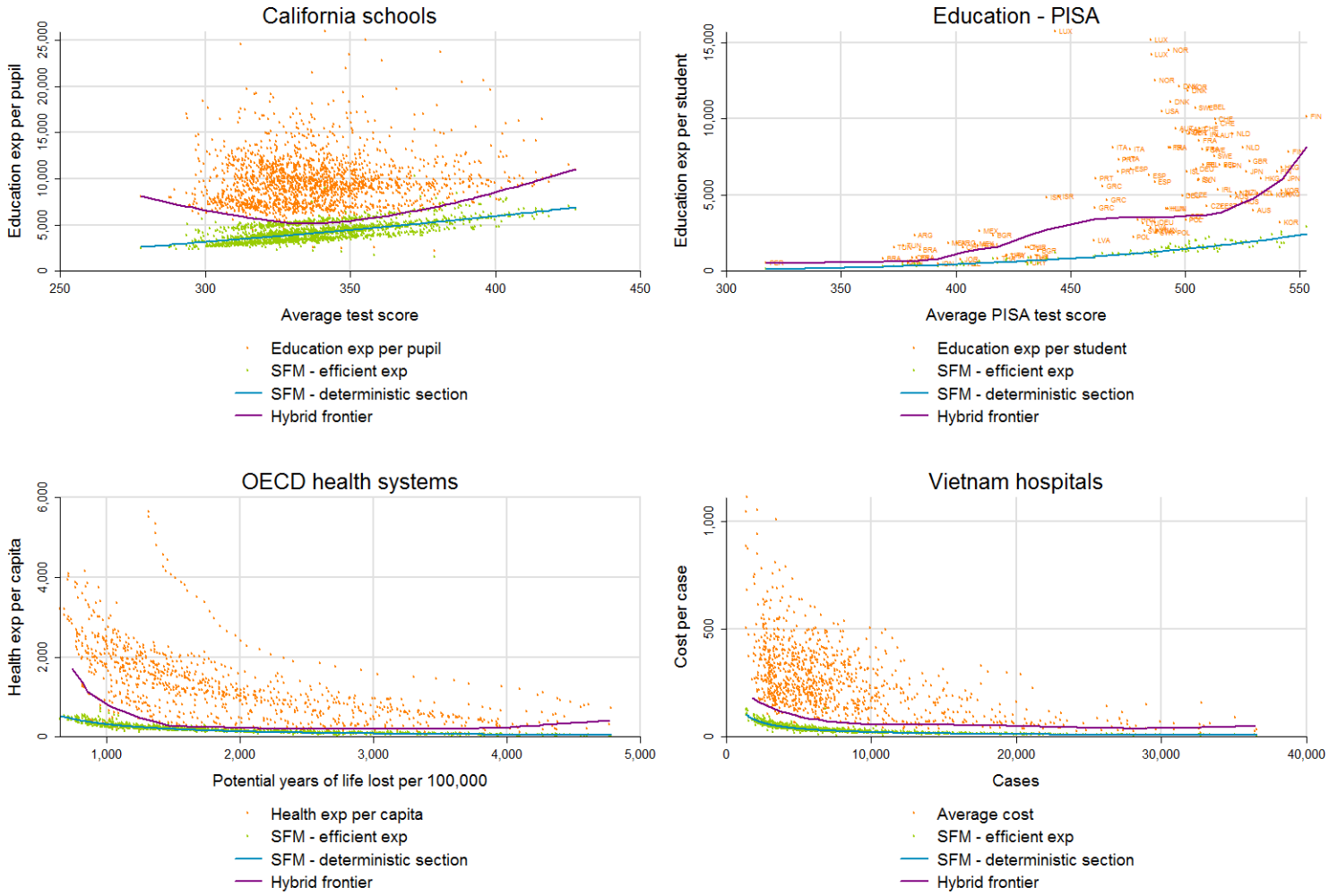
Notes: California data refer to 313 unified school districts, and cover the period 2003/2004 - 2008/2009. The PISA data refer to 49 countries, and cover some or all the years 2000, 2003 and 2006. The OECD health system data refer to 29 countries, and cover the period 1960-2005. The Vietnam data refer to 795 district hospitals with between 50 and 500 beds, and cover the period 1998-2000.

Figure 2: Least efficient units in the four examples



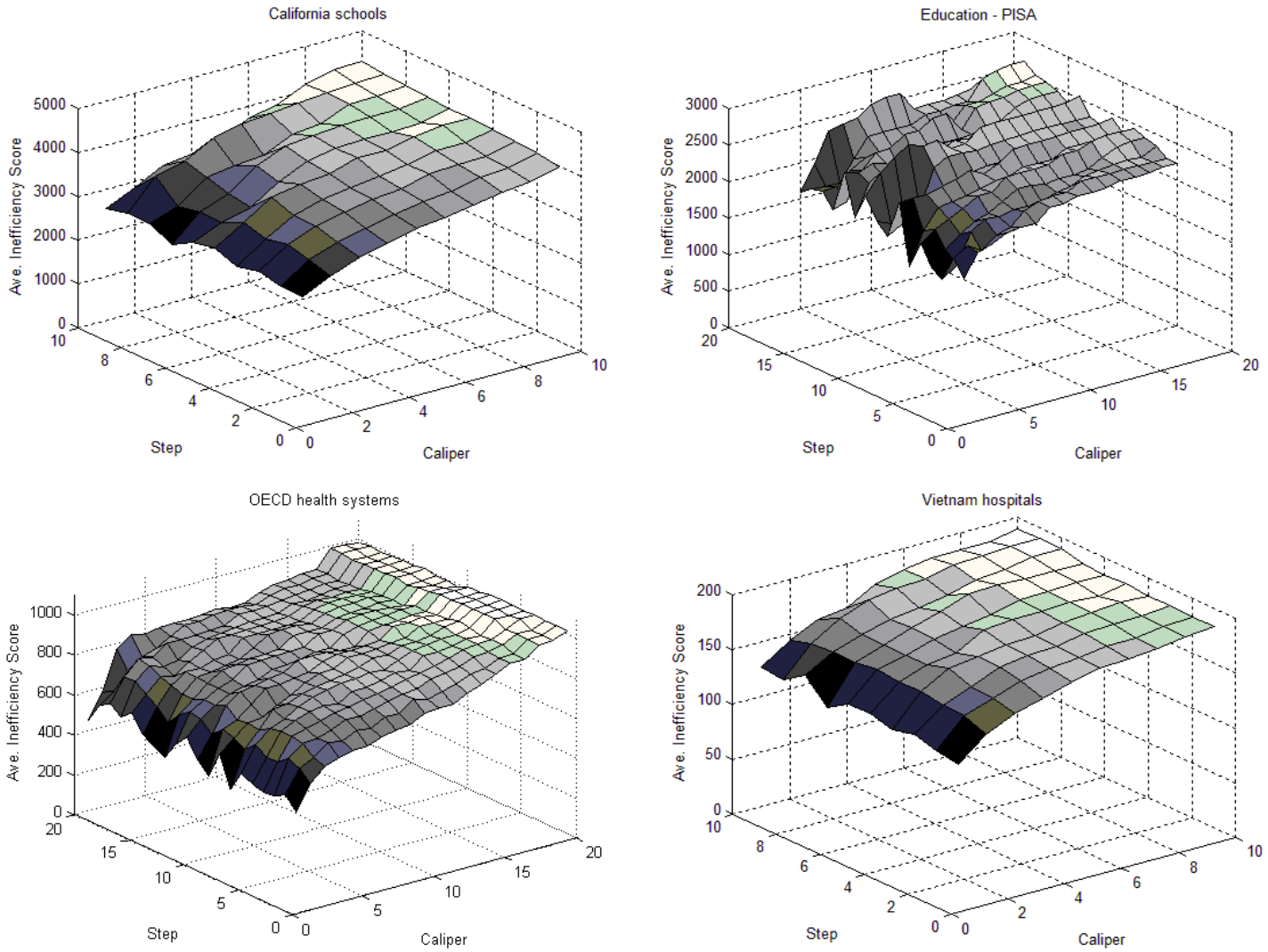
Notes: California data refer to 313 unified school districts, and cover the period 2003/2004 - 2008/2009. The PISA data refer to 49 countries, and cover some or all the years 2000, 2003 and 2006. The OECD health system data refer to 29 countries, and cover the period 1960-2005. The Vietnam data refer to 795 district hospitals with between 50 and 500 beds, and cover the period 1998-2000. Units are averaged over the sample period. Vietnam hospital codes refer to region and hospital identifier.

Figure 3: Comparison of hybrid and stochastic frontier methods



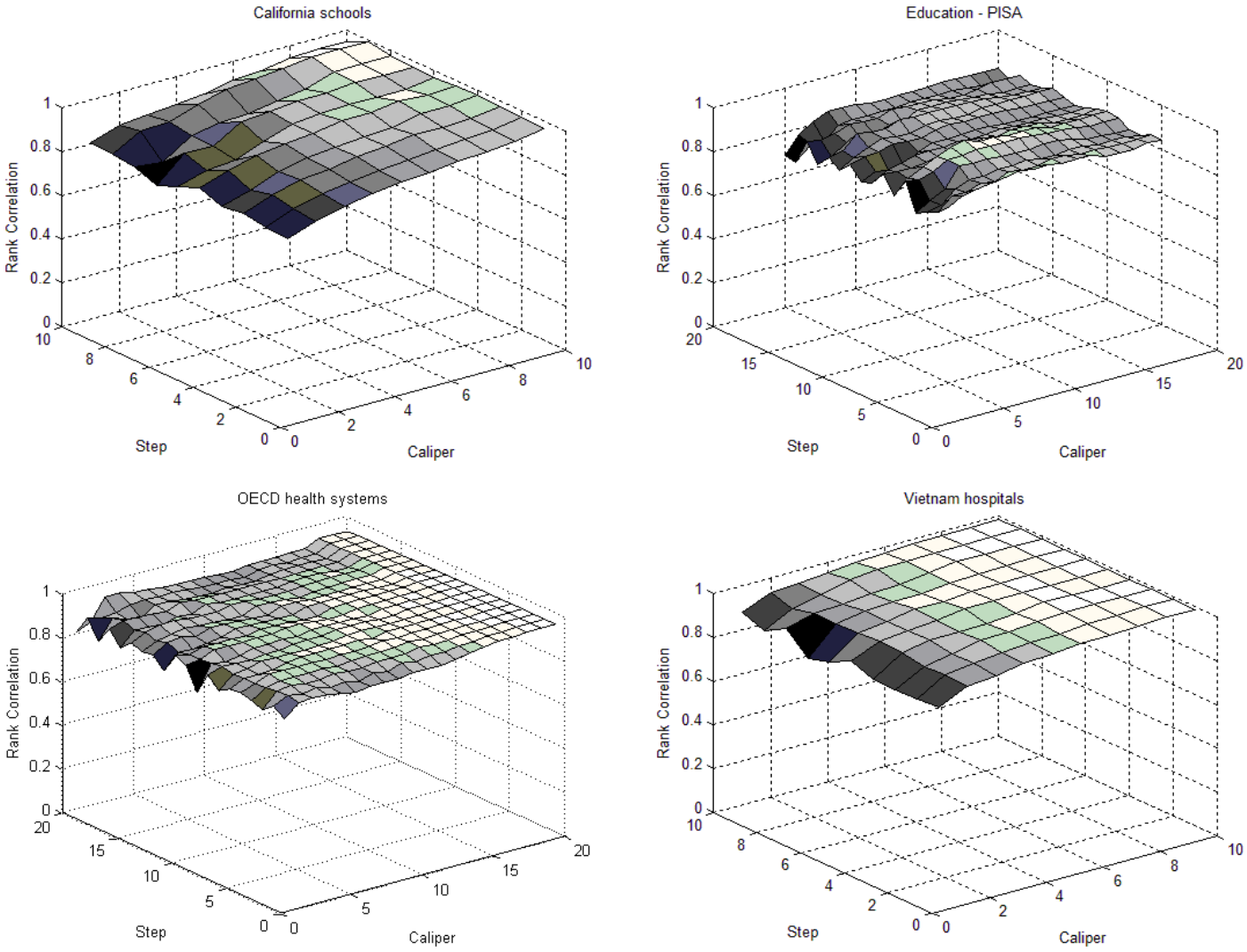
Notes: Stochastic frontier model is time-invariant half-normal panel-data model. California data refer to 313 unified school districts, and cover the period 2003/2004 - 2008/2009. The PISA data refer to 49 countries, and cover some or all the years 2000, 2003 and 2006. The OECD health system data refer to 29 countries, and cover the period 1960-2005. The Vietnam data refer to 795 district hospitals with between 50 and 500 beds, and cover the period 1998-2000.

Figure 4: Sensitivity of single-output inefficiency estimates to choice of caliper and step



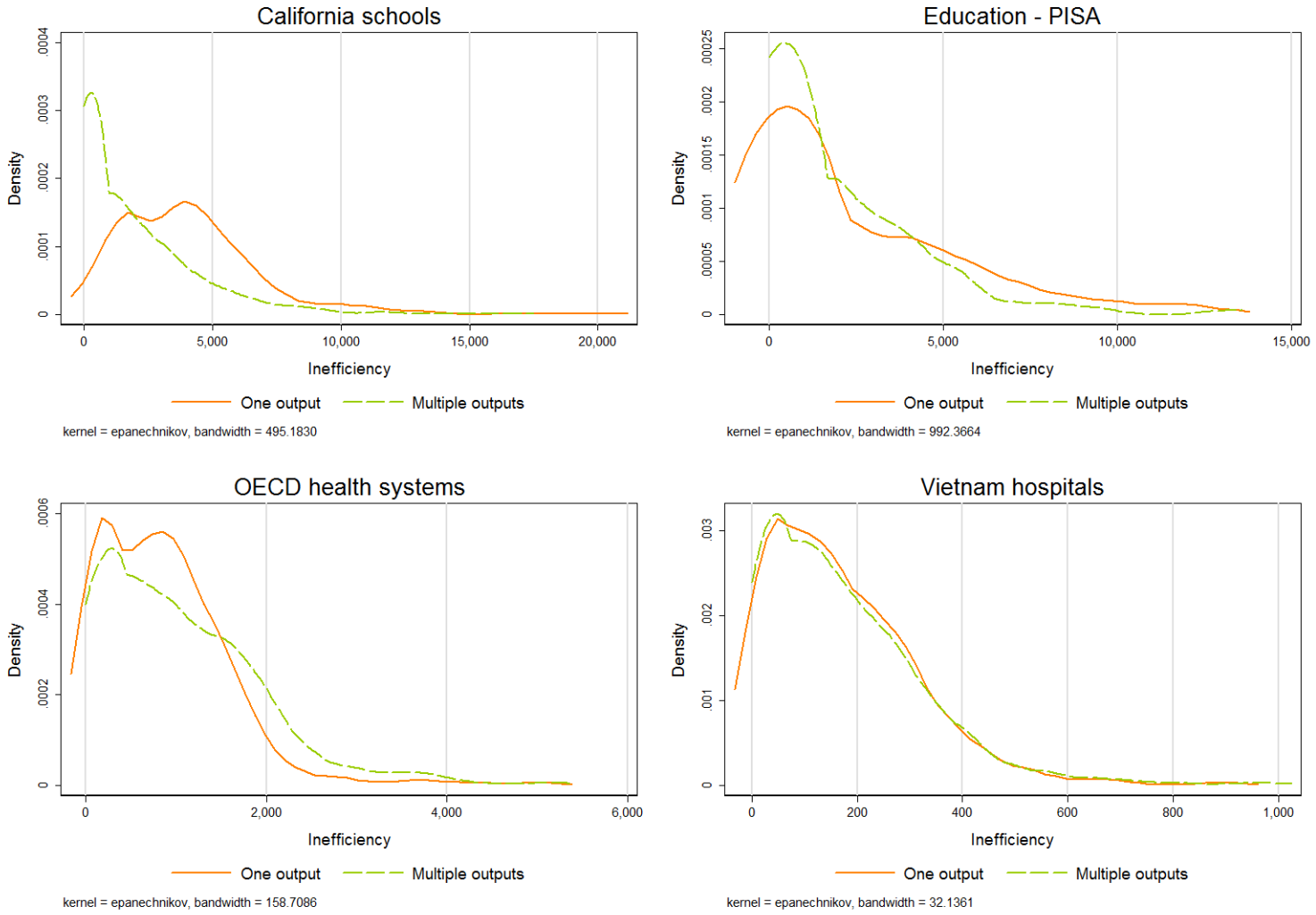
Notes: Charts show average inefficiency score among non-frontier units obtained with caliper and step combination shown. Actual caliper and step used are 1/100 of the values shown. Caliper and step are defined in terms of standard deviation of the outcome variable.

Figure 5: Sensitivity of single-output rankings to choice of caliper and step



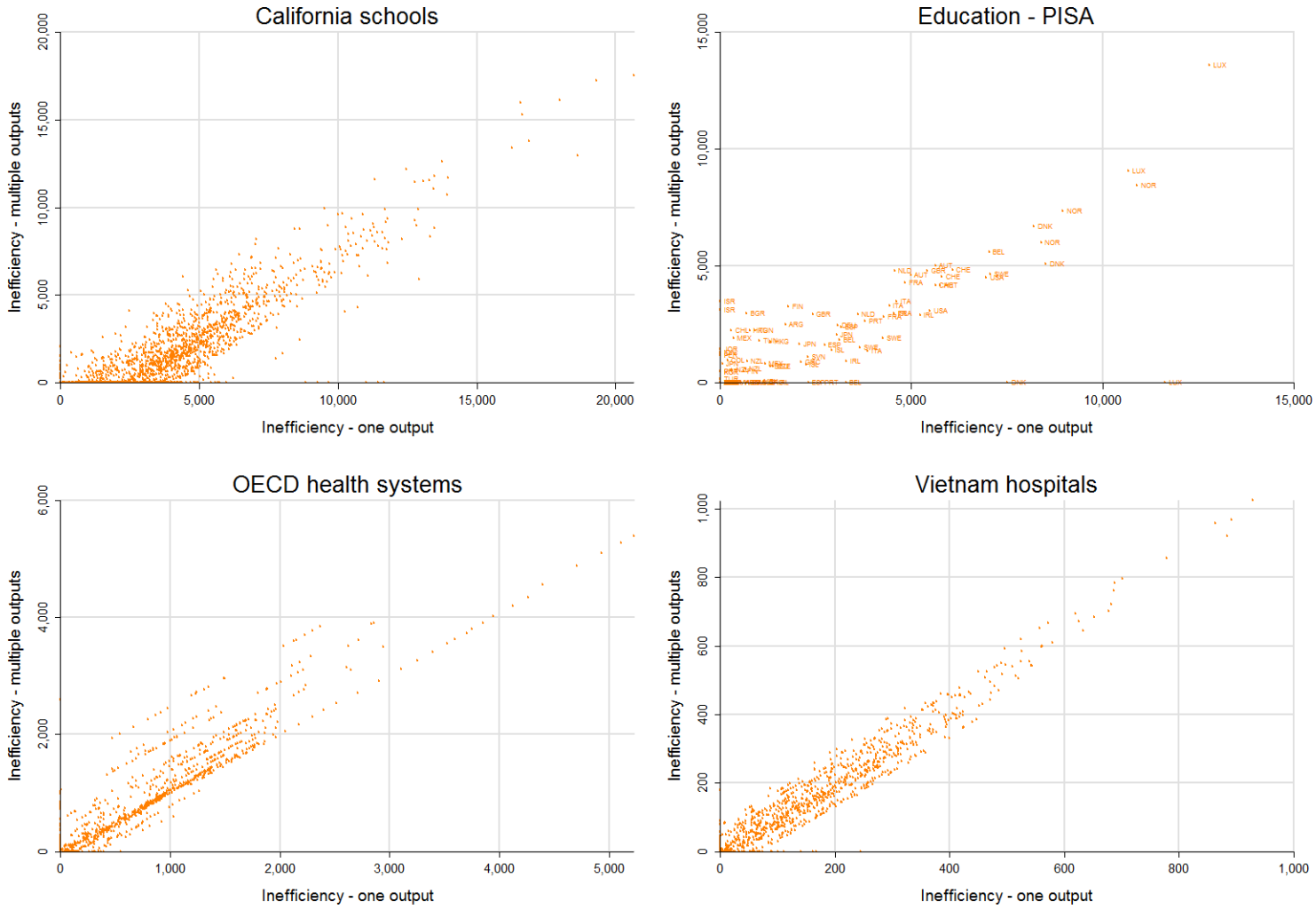
Notes: Charts show rank correlation between inefficiency score obtained with caliper and step combination shown, and inefficiency score with caliper and step combination used in basic analysis, as reported in Table 2. Actual caliper and step used are 1/100 of the values shown. Caliper and step are defined in terms of standard deviation of the outcome variable.

Figure 6: Comparison of single- and multiple-output efficiency frequency distributions



Notes: California data refer to 313 unified school districts, and cover the period 2003/2004 - 2008/2009. The PISA data refer to 49 countries, and cover some or all the years 2000, 2003 and 2006. The OECD health system data refer to 29 countries, and cover the period 1960-2005. The Vietnam data refer to 795 district hospitals with between 50 and 500 beds, and cover the period 1998-2000.

Figure 7: Single- vs. multiple-output results



Notes: California data refer to 313 unified school districts, and cover the period 2003/2004 - 2008/2009. The PISA data refer to 49 countries, and cover some or all the years 2000, 2003 and 2006. The OECD health system data refer to 29 countries, and cover the period 1960-2005. The Vietnam data refer to 795 district hospitals with between 50 and 500 beds, and cover the period 1998-2000.

Figure 8: Allowing poverty rates to be an exogenous influence on California's school districts' efficiency

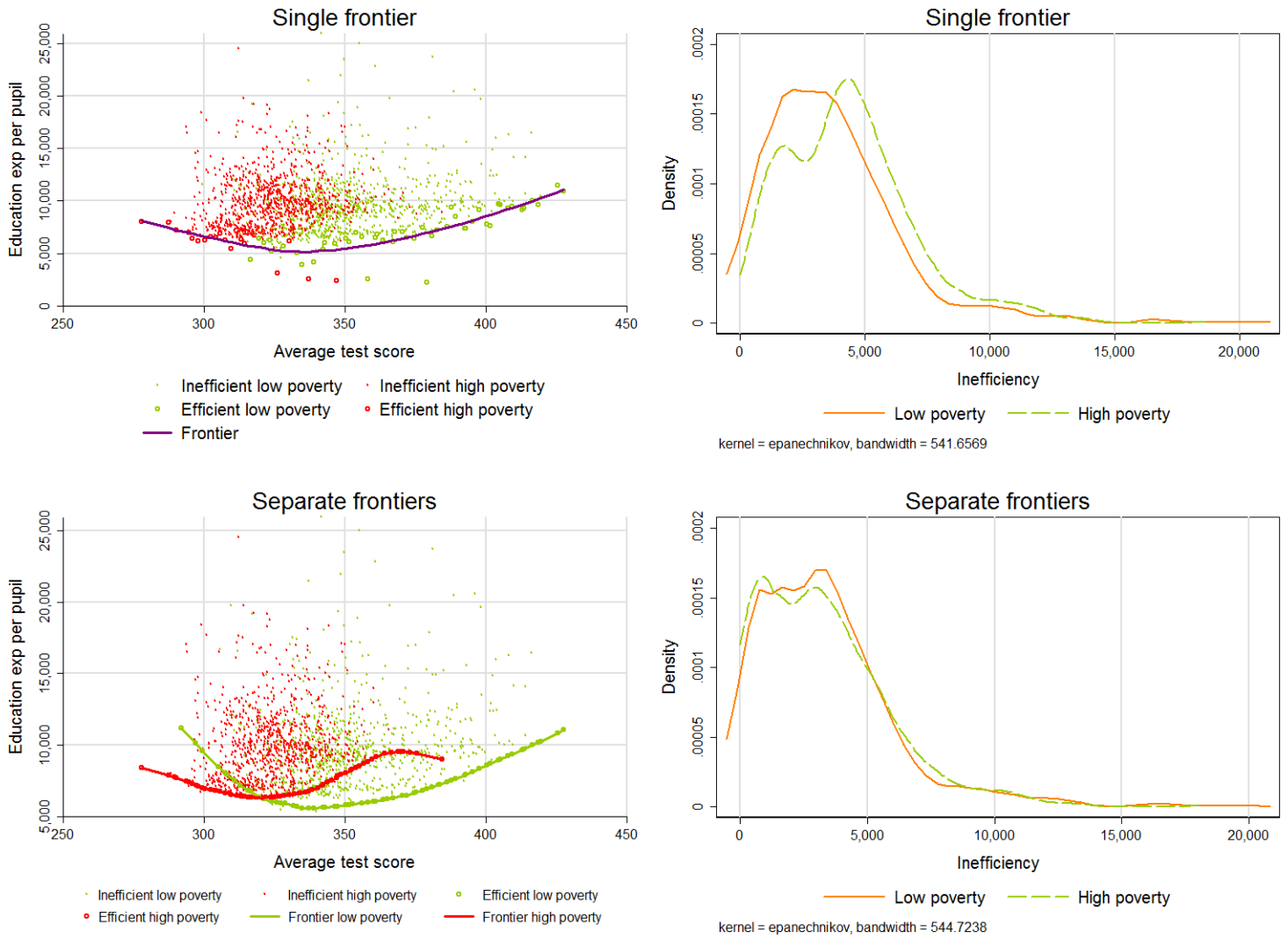
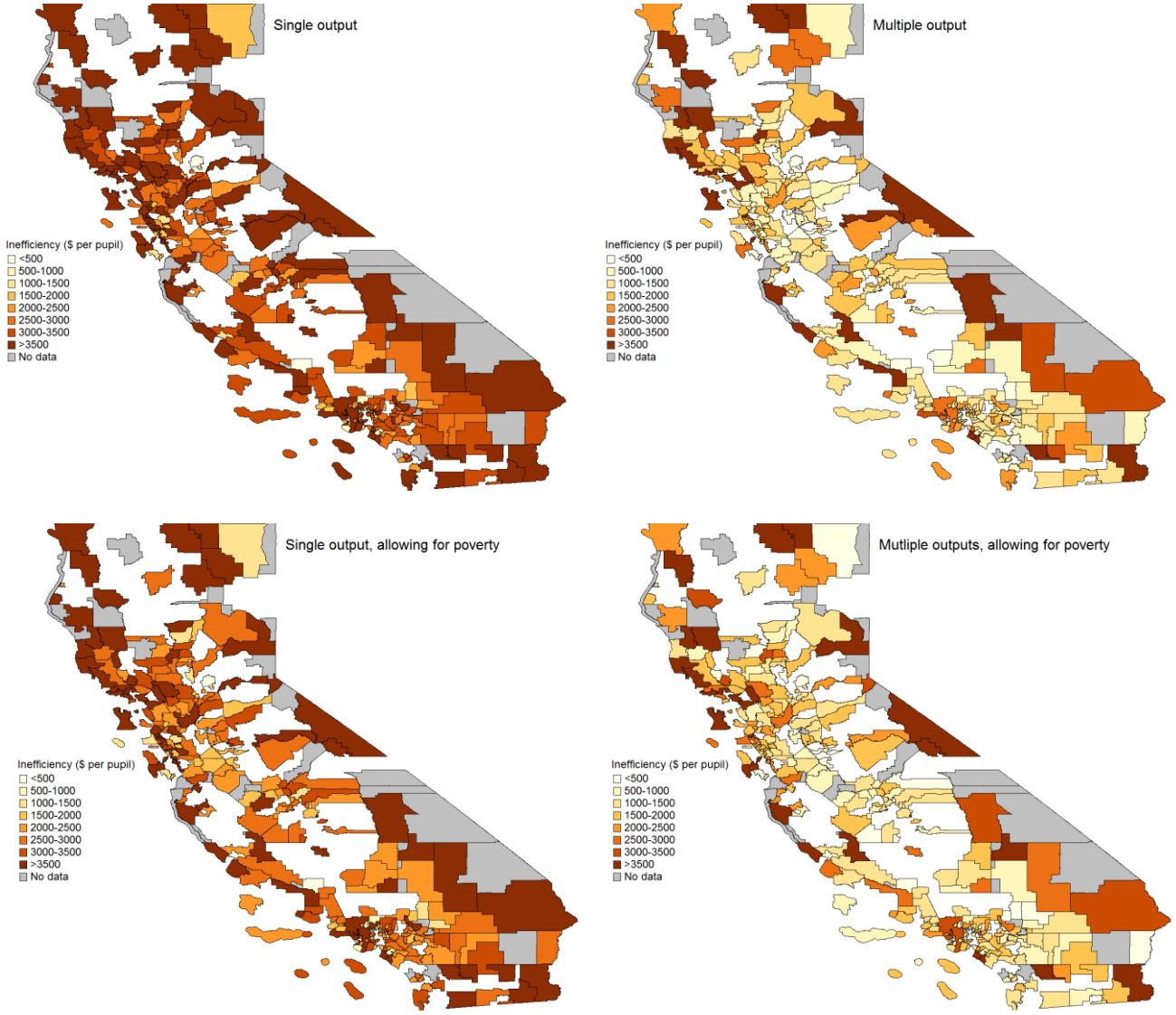


Figure 9: California's school districts' inefficiency



References

Afonso, A. and M. St Aubyn (2005). "Nonparametric approaches to education and health efficiency in OECD countries." Journal of Applied Economics **8**(2): 227-246.

Ammermueller, A. (2007). "PISA: What makes the difference?" Empirical Economics **33**(2): 263-287.

Anderson, G. F. and B. K. Frogner (2008). "Health spending in OECD countries: obtaining value per dollar." Health Aff (Millwood) **27**(6): 1718-27.

Bessent, A. M. and E. W. Bessent (1980). "Determining the comparative efficiency of schools through data envelopment analysis." Educational Administration Quarterly **16**(2): 57.

Burgess, J. F. (2006). Productivity Analysis in Health Care. In: The Elgar Companion to Health Economics(ed). Cheltenham, U.K. and Northampton, Mass., Elgar**:** 335-42.

Chambers, J., J. Levin and D. DeLancey (2006). "Efficiency and Adequacy in California School Finance: A Professional Judgment Approach." American Institutes for Research: 88.

Coelli, T. (1996). A guide to DEAP version 2.1: a data envelopment analysis (computer) program. Worthington, Centre for Efficiency and Productivity Analysis, University of New England, Australia.

Costrell, R., E. Hanushek and S. Loeb (2008). "What Do Cost Functions Tell Us About the Cost of an Adequate Education?" Peabody Journal of Education **83**(2): 198-223.

Crampes, C. and A. Hollander (1995). "Duopoly and Quality Standards." European Economic Review **39 1**: 71-82.

Deller, S. C. and E. Rudnicki (1993). "Production Efficiency in Elementary Education: The Case of Maine Public Schools." Economics of Education Review **12 1**: 45-57.

Döbert, H., E. Klieme and W. Sroka (2004). Conditions of School Performance in Seven Countries: a quest for understanding the international variation of PISA results, Waxmann Verlag.

Freedberg, L. and S. K. Doig (2011). Spending far from equal among state's school districts, analysis finds. California Watch.

Fuchs, T. and L. Woessmann (2007). "What accounts for international differences in student performance? A re-examination using PISA data." Empirical Economics **32**(2): 433-464.

Hakkinen, U. and I. Joumard (2007). Cross-country Analysis of Efficiency in OECD Health Care Sectors: Options for Research. OECD Economics Department, OECD Economics Department Working Papers, 554.

Hollingsworth, B. (2008). "The Measurement of Efficiency and Productivity of Health Care Delivery." Health Economics **17 10**: 1107-28.

Hollingsworth, B. and A. Street (2006). "The Market for Efficiency Analysis of Health Care Organisations." Health Economics **15 10**: 1055-59.

Huang, Y. G. and C. P. Mclaughlin (1989). "Relative Efficiency in Rural Primary Health-Care - an Application of Data Envelopment Analysis." Health Services Research **24**(2): 143-158.

Imazeki, J. (2006). Assessing the Costs of K-12 Education in California Public School, Governor's Committee on Education Excellence.

Leuven, E. and B. Sianesi (2003). PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing, Boston College Department of Economics.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. In: Proceedings of the National Institute of Science(ed). Calcutta. **12:** 49.

Nolte, E. and C. M. McKee (2008). "Measuring the health of nations: updating an earlier analysis." Health Aff (Millwood) **27**(1): 58-71.

Ray, S. C. (1991). "Resource-Use Efficiency in Public-Schools - a Study of Connecticut Data." Management Science **37**(12): 1620-1628.

Schmidt, P. and R. C. Sickles (1984). "Production Frontiers and Panel Data." Journal of Business and Economic Statistics **2 4**: 367-74.

Sherman, H. D. (1984). "Hospital efficiency measurement and evaluation: empirical test of a new technique." Medical Care **22**(10): 922-938.

Simar, L. and P. W. Wilson (2000). "Statistical inference in nonparametric frontier models: The state of the art." Journal of Productivity Analysis **13**(1): 49-78.

Wagstaff, A. (1989). "Estimating Efficiency in the Hospital Sector - a Comparison of 3 Statistical Cost Frontier Models." Applied Economics **21**(5): 659-672.

Weaver, M. and A. Deolalikar (2004). "Economies of scale and scope in Vietnamese hospitals." Soc Sci Med **59**(1): 199-208.

Worthington, A. C. (2001). "An empirical survey of frontier efficiency measurement techniques in education." Education Economics **9**(3): 245-268.